Speech Emotion Recognition Based on Fusion Method

Sara Motamed
Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
samotamed@yahoo.com
Saeed Setayeshi*
Department of Medical Radiation, Amirkabir University of Technology, Tehran, Iran
setayesh@aut.ac.ir
Azam Rabiee
Department of Computer Science, Dolatabad Branch, Islamic Azad University, Isfahan, Iran
azrabiee@gmail.com
Arash Sharifi
Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
a.sharifi@srbiau.ac.ir

Received: 27/Jul/2016 Revised: 07/Jan/2017 Accepted: 14/Jan/2017

Abstract

Speech emotion signals are the quickest and most neutral method in individuals' relationships, leading researchers to develop speech emotion signal as a quick and efficient technique to communicate between man and machine. This paper introduces a new classification method using multi-constraints partitioning approach on emotional speech signals. To classify the rate of speech emotion signals, the features vectors are extracted using Mel frequency Cepstrum coefficient (MFCC) and auto correlation function coefficient (ACFC) and a combination of these two models. This study found the way that features' number and fusion method can impress in the rate of emotional speech recognition. The proposed model has been compared with MLP model of recognition. Results revealed that the proposed algorithm has a powerful capability to identify and explore human emotion.

Keywords: Speech Emotion Recognition; Mel Frequency Cepstral Coefficient (MFCC); Fixed and Variable Structures Stochastic Automata; Multi-constraint; Fusion Method.

1. Introduction

Emotion recognition refers to the ability of detecting humans' feelings and conditions. It is also one of the most efficient methods of analyzing information collected from humans to identify the interaction between man and machines [1,2]. Most researches of recognition have focused either on its recognition or classification [3]. Seehapoch et al. have claimed that there are two elements in the speaker feeling including prosodic and spectral which may influence on speech emotion recognition since both of them contain emotional information [4]. Many researchers have tried to separate the speech features including pitch, energy, frequency, formant and vibration [5].

There are many debates regarding identification of speech emotion recognition about having insufficient knowledge for identification of speech sounds, lack of powerful database and even different accents and dialect expression. Hozjan et al. have assumed that there should be no difference in the culture of the speakers, so they ignored cultural changes in their behavior in analyzing problems associated with emotional speech recognition [6]. Cahn proved that acoustic features including intonation, sound quality and enumeration have powerful relationship with speech recognition [7].

To increase the rate of recognition, some researchers combined the extracted features for speech recognition. Zhang et al. used the Gaussian mixture model (GMM), Mel frequency Cepstral coefficient (MFCC) and autocorrelation function coefficients (ACFC) in the feature extraction level. Their results indicated a suitable rate of recognition [8]. Similarly, Bojanic et al. employed a fusion method of speech emotion recognition to identify the speaker feelings through the use of neural network method [9]. Since identification of speech recognition is an important problem to extract speech meaning, this paper uses a special fusion method to classify emotional speech through using the parallel stochastic learning automata.

This paper introduces a new classification model which employs multi-constraints with the use of stochastic learning automata to recognize speech emotion. The rest of the paper is organized as follows. Section 2 discusses about feature extraction methods. The proposed method and stochastic learning automata are explained in sections 3 and 4, respectively. The test results and conclusion are given in sections 5 and 6, respectively.

2. Features Extraction

2.1 Mel Frequency Cepstrum Coefficients (MFCC)

One of the most important steps in speech emotion recognition is to extract suitable features. Speech features are divided into the four main groups and numerous

* Corresponding Author

methods are introduced by researchers to obtain a powerful feature extractor. These four groups include speech continuous, qualitative, spectral and TEO based classes [4]. Price worked on the speech sound and energy features and grouped these under spectral classes [10]. MFCC algorithm is one of the efficient methods grouped under spectral features [10].

Many useful features can be extracted from the speech signals such as energy, MFCC (Mel frequency cepstral coefficients), LPC (linear predictive coding) and so on. This set of features has important information for discriminating different types of emotions [11] [12]. In this work, we have selected the MFCC to extract the emotional features [2,13]. This method includes the following mathematical approaches:

Step 1- Pre-emphasize:

The signal passes a filter of high frequency emphasized by equation (1) [14].

$$Y[n] = X[n] - 0.95\,X[n-1] \qquad (1)$$

Step 2- Framing:

To classify speech samples, analogue conversations are changed into digital conversations in small frames length of 20 to 40 ms.

The speech signal is divided in N frames. Vicinity of separated frames by M is (M<N) where M=100 and N=256 [14].

Step 3- Hamming Window:

Windows are defined with W(n) and $0 \leq n \leq N\text{-}1$. N is the number of samples in each frame and Y[n] is the output signal. Input signals are shown by X(n) and windowing results are shown in equations (2) and (3) [14].

$$Y(n) = X(n) \times W(n) \qquad (2)$$

$$w(n) = 0.54 - 0.4\cos\left[\frac{2\Pi n}{N} - 1\right]$$
$$0 \leq n \leq N-1 \qquad (3)$$

Step 4- Fast Fourier Transform:

Each frame is changed into frequency amplitude from time amplitude with N. The Fourier Transform is used to reverse pulse complexity related to glottal U[n] and instigation of vocal tract H[n] in time domain and is calculated by equation (4), [14].

$$y(w) = \mathrm{FFT}\,[h(t) * X(t)] = H(w) * X(w) \qquad (4)$$

Step 5- Processing Mel Filter bank:

Cepstral coefficients are mainly obtained from the output of a set of filter banks, which suitably cover the full range of defined frequency spectrum. Generally the set of 24 buffer filters were used, which are similar to human ear performance. Filters were placed along the frequency axes variably. More filters were allocated to part of the spectrum below 1 KHz since they have more information about sound. Filtering is referred to as conceptual weighting.

Filter outputs are a set of their filtered spectral components: equation (5) indicates calculation of Mel for given frequencies in Hz [14]:

$$F(\mathrm{mel}) = [2595 * \log 10\,[1 + f]700] \qquad (5)$$

Step 6- Cos transform:

This step is a process of transferring Mel spectrum to time span using discrete Cosine transform (DCT). Results are Cepstral coefficients for Mel frequency and the set of coefficients are referred to as acoustic vectors [14]. The obtained features vector is put into the next step for partitioning. The block diagram for our recommended model is shown in fig 1.

2.2 Auto Correlation Function Coefficients (ACFC)

Auto Correlation Function of periodic signal generates a maximal value when the delay equals to the function cycle [8]. So, this method can find the maximum value pith period of the signal. The autocorrelation function calculated by equation 6:

$$p(k) = \lim_{n\to\infty} \frac{1}{2N+1} \sum_{-\infty}^{+\infty} x(m)x(m+k) \qquad (6)$$

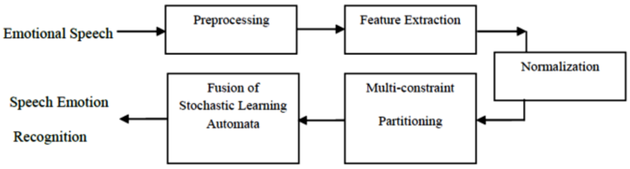
Emotional Speech
Preprocessing
Feature Extraction
Normalization
Multi-constraint Partitioning
Fusion of Stochastic Learning Automata
Speech Emotion Recognition

Fig. 1 Recommended system block diagram

Unvoiced signal and its autocorrelation function have no periodicity, and no significant peak. P (k) rapidly decays while k increases. Voiced signal has a quasi-periodicity, and its auto-correlation function P (k) has the same period with k [8].

3. Multi- constraint Partitioning

In order to classify obtained features, the MFCC features are used to allocate collected features of speech emotion recognition signals to the groups with the highest similarity. The proposed model is an efficient approach to allocate P features vectors values obtained by MFCC, with the |P| elements, to N classes (N is the number of people in our database) where each class has Ni Nodes with the certain capacity (Ni the number of different actions of each person). This means that each set of feature vectors has limited capacity considering constraints on the connections. External and internal equations to explore limitations and relationships are presented here.

Xi,n is an index which has a value of either 0 or 1 ($X_{i,n}\ \epsilon\{0,1\}$) and if features vector P_i is allocated to N_n node, then its quantity would be 1 otherwise it would be 0 [12]. External relationship is said to be a node with the supposition that P_i has been allocated to this specific Node [12]. Then the external connection of P_i process is then applied to all feature vectors for P_j, $\sum_{j=1}^{|P|}\left(1-X_{i,n}\right)w_{i,j}$. If each P_j process is allocated to N_n node then $X_{j,n}=1$ and this process has no participation with the above sum. Equation 7 calculates the external relationship of the nodes [12].

$$\sum_{i=1}^{|P|}\sum_{j=1}^{|P|}\left(X_{i,n}-X_{j,n}X_{j,n}\right)W_{i,j}\leq 1$$
$$n=1,\dots,|N| \qquad (7)$$

The above formula divides features vectors sets into the subsets and adds the connection from one set to another. The only guiding quantity is that of $w_{i,j}$ connection which indicates connection from P_i to the node P_j (remote features vectors which is not connected to the node). Such internal connection may therefore be obtained by a similar formula. However, by adding the connection to $w_{j,i}$ feature vector to the node from remote feature vector the result of which would be equation (8), [12].

$$\sum_{i=1}^{|P|}\sum_{j=1}^{|P|}\left(X_{i,n}-X_{j,n}X_{j,n}\right)W_{j,i}\leq 1$$
$$n=1,\dots,|N| \qquad (8)$$

The set of limitations in equation (8), limits the total calculation time for feature vectors allocated to each node with normalized capacity of node. The set of limitations in equation (9) assures that each feature vector would be placed on one node only [12].

$$\sum_{i=1}^{|P|}X_{i,n}\square_i \leq 1$$
$$n=1,\dots,|N| \qquad (9)$$

$$\sum_{i=1}^{|N|}X_{i,n} \leq 1$$
$$n=1,\dots,|P \qquad (10)$$

4. Stochastic Learning Automata

4.1 Fixed Structure Stochastic Automata (FSSA)

In an identification cycle, an automaton would select a behavior considering the reward and penalty received from the environment [15]. The automaton then uses the collected answer and the knowledge of former behavior toward defining the next measure. The objective of learning automaton is to perform an optimized measure beyond permitted behaviors. The automaton matches with the environment by learning optimal operation. The node with the most trespass from equation 7 – external connection for all vectors- is selected for pairing and computing of w_i. A feature vector, for instance P_A elected randomly between feature vectors on the node considering experimental distribution from their τ_i weight, is allocated to this node. The P_A feature vector randomly selects another P_B feature vector According to W_A distribution probability. The total feature vectors $\langle P_A, P_B\rangle$ are considered as a successful pair. If two feature vectors belong to the same node then those two feature vectors would receive a reward, unless their pairs are unsuccessful in which case both are penalized [15].

Learning samples modeled by learner automaton applications haves been found in systems with insufficient knowledge about the environment and their startup. FSSA output functions and transfer do not change over time. The problem is based on the fact that a stable map of a subclass is obtained from learning automaton solutions and is used for solving object partitioning problems.

4.2 Variable Structure Stochastic Automata (VSSA)

VSSA are a replacement for FSSA and their transfer and output matrixes are changed in time [15,16,17]. VSSA are defined as a possible operation vector P(K) where P_i (K) is the possibility of ith operation in the set of A operation which is selected in K time from available |A| operations. As $\sum_i P_i\,(K)=1$ for all Ks, updating the law for possibility vectors would be continuous or discontinuous. The VSSA family has the quickest learning automaton convergence. Combining the VSSA family with automata for solving the specified problems will hopefully improve the speed of obtaining a solution [15,16,17]. The main characteristic of estimating algorithms is that they maintain estimations of possible rewards from each operation and use them in updating probability equations. In essence, an automaton selects an operation in the first step of the function cycle and produces the environment for answering the operation. Estimation of possible rewards for that operation is updated according to the estimating algorithm answer.

4.2.1 Discrete Generalized Pursuit Learning Automata (DGPA)

One of the problems of standard learning algorithms is their relatively slow convergence in selecting the optimum operation in static environments where several solutions have been introduced. These solutions are based on the disconnection of possibility space, in which the possibility of operation selection can pick only certain quantities in the range of [0,1] [15,16,17]. An existing problem in new models is premature convergence of learning algorithms with non-optimized operations, which is rooted in the limitation of possibility space. To improve learning algorithms convergence, Thathachar and Sastry introduced a new route in estimator algorithms. The most important specification of these algorithms is maintaining a continuous estimation based on the possibility of receiving the reward for any operation, and using it to update automata equations. In other words, in the first step of an operation cycle, automata would select an operation and then produces an environmental response for it. According to this response, the estimating algorithm would present reward possibility estimation for that operation. One group of estimating algorithms is called pursuit algorithms.

Pursuit algorithms are divided in two classes of continuous and discrete [18]. The difference between these two classes lies in updating the law for operation possibilities. According to results in [18], partitioning on the basis of pursuit automata with variable structure would only work for small classes. Therefore, this paper have been used the fusion of parallel discrete pursuit learning automata, that is defined in section 4-2-2, to solve multi-constraint problems.

4.2.2 Fusion of DGPA

Learning rate and convergence speed equivalence are important paramters in learning automata. The parallel operations method is considered to increase convergence speed in environment.

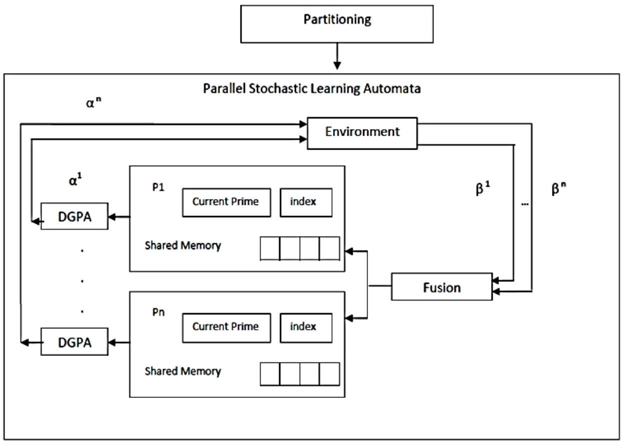

Fig. 2. Fusion of stochastic learning automata

Our proposed model is considered to be parallel instead of single learning automaton. Fig. 2 presents the allocation of |P| feature vector to |N| classes where next limitations are applied. At this stage, speech signals are divided into n parts and given to processors that the learning of discrete pursuit automata is applied to them in a parallel manner. In fig.2, n parameters are a total number of n={1,…,n} operations and P is the number of processors where each processor in this case is an automaton. “Indexes” in our model are stable and “Current Primes” in each moment are variable. Input vectors are divided by each processor to a certain number. Processors output is the input of DGPA.

In the fusion section, each set of operations is given by α, while the operation possibility vector P(K) is common to all n automata. Each instance of discrete pursuit learning is based on the common operation possibility vector, selected from (α^i(k)). This vector obtains its own reinforcement signal ((β^i(k)). It is supposed that β^j(k)ϵ [0,1] is obtained for all i and k and the fusion vector are computed by the equation (11) [18].

$$q_i(k) = \sum_{j=1}^{n} \beta^j\,(k) I\{\alpha^j(k) = \alpha_i\} \qquad (11)$$

Equation (12) calculates the total response by simple summation:

$$q(k) = \sum_{j=1}^{n} \beta^j\,(k) = \sum_{j=1}^{n} q_i\ \ (k) \qquad (12)$$

The output of fusion step is obtained by equation (13):

$$p_i(k+1) = p_i(k) + \tilde{\square}\,\big(\,q_i(k) - q(k)p_i(k)\big)$$
$$,i = 1, \ldots, r \qquad (13)$$

λ ϵ (0,1) is learning parameter and $\tilde{\lambda}\ = {}^{\lambda}/_{n}$ is its normalized value. Considering computations in equation (13), P(K) has been updated only once. The updated value of p(k+1) is shared by all automata for selection of the next operation. This algorithm is suitable for n sizes and speeding up the rate of convergence.

5. Test Result and Conclusion

5.1 Dataset

The “Berlin Dataset of Emotional Speech” was used to train and test the algorithm in this paper [19,20]. The Berlin Dataset consists of 535 speech samples, consisting of German utterances related to emotions such as anger, disgust, fear, happiness, sadness, surprise, and neutrality, as performed by five male and five female voice actors. Each one of the ten professional actors expresses ten words and five sentences covering each of the emotional categories. The corpus was evaluated by 25 judges who classified each emotion with a score rate of 80%.

This Dataset was chosen for the following reasons: (i) the quality of its recording is very good and (ii) it is a public and popular Dataset of emotion recognition that is recommended in the literature [21]. This paper has used on the Berlin Dataset in order to achieve a higher and more accurate rate of recognition.

5.2 Lab Results

This paper introduces a new method of classification for speech emotion recognition. Signals are normalized and then fed to the MFCC and ACFC. After feature extraction by MFCC and ACFC, all features vectors are sent to the proposed classification model. Tables 1 to 7 indicate the experimental results on emotional speech signals.

Table 1. Speech emotion recognition using MFCC and single FSSA

FSSA	ANGER	JOY	NEUTRAL	DISGUST	FEAR	SADNESS
ANGER	**85.0**	10.0	10.00	15.0	10.0	0.0
JOY	0.0	**15.0**	0.20	0.0	0.0	0.0
NEUTRAL	15.0	65.0	**89.80**	30.0	55.5	40.0
DISGUST	0.0	0.0	0.0	**25.0**	0.5	10.0
FEAR	0.0	10.0	0.0	5.0	**20.0**	22.0
SADNESS	0.0	0.0	0.0	25.0	10.0	**28.0**

Table 1 shows that the highest rate of emotional learning is 89.80 percent by using MFCC and FSSA learning models for the neutral state and 85% for the state of anger.

Table 2. Speech emotion recognition using ACFC single FSSA

FSSA	ANGER	JOY	NEUTRAL	DISGUST	FEAR	SADNESS
ANGER	**83.25**	8.0	9.0	10.0	10.0	4.0
JOY	0.0	**10.0**	5.0	8.0	0.01	1.0
NEUTRAL	6.75	72.0	**86.0**	25.0	45.00	35.0
DISGUST	0.0	0.0	0.0	**20.0**	5.0	10.0
FEAR	0.0	10.0	0.0	9.0	**20.0**	21.0
SADNESS	0.0	0.0	0.0	28.0	10.0	**29.0**

Table 2 shows that the rate of emotional learning is 86.00% by using ACFC and FSSA learning models for the neutral state and 83.25% for the state of anger. By comparison of table 1 and table 2, the highest rate of recognition belongs to MFCC and FSSA.

To improve the rate of recognition, the fusion of MFCC and ACFC have been employed. Table 3 presents the rates of speech emotion recognition by using the fusion of MFCC, ACFC and FSSA classification.

Table 3. Speech emotion recognition using MFCC and ACFC with single FSSA

FSSA	ANGER	JOY	NEUTRAL	DISGUST	FEAR	SADNESS
ANGER	86.0	8.5	9.0	19.0	10.0	0.0
JOY	1.0	15.0	0.0	0.0	0.0	0.0
NEUTRAL	13.0	67.00	91.0	25.0	40.0	44.5
DISGUST	0.0	0.0	0.0	30.5	5.0	5.5
FEAR	0.0	9.5	0.0	0.5	30.0	10.0
SADNESS	0.0	0.0	0.0	25.0	15.0	40.0

According to table 3, the highest and lowest rate of recognition belong to neutral and joy, respectively. Table 4 indicates the rate of emotional speech recognition using the combination of MFCC, ACFC and fusion of FSSAs learning model.

Table 4. Speech emotion recognition using the combination of MFCC, ACFC and fusion of FSSAs

Fusion of FSSAs	ANGER	JOY	NEUTRAL	DISGUST	FEAR	SADNESS
ANGER	**88.50**	10.0	7.30	10.0	15.0	0.0
JOY	0.0	**17.0**	0.0	0.0	0.0	0.0
NEUTRAL	11.50	62.5	**92.70**	25.0	67.5	58.0
DISGUST	0.0	0.0	0.0	**30.0**	0.5	0.0
FEAR	0.0	10.5	0.0	1.0	**17.0**	22.0
SADNESS	0.0	0.0	0.0	25.0	0.0	**20.0**

According to table 4, the highest rate of recognition belongs to neutral state by 92.70% followed by anger with the value of 88.50%. Therefore, it can be claimed that FSSAs fusion algorithm gives better results than FSSA single model.

Table 5 shows the rate emotional speech recognition by using the combination of MFCC, ACFC and fusion of DGPAs classification method with two parallel processors.

Table 5. Speech emotion recognition using the combination of MFCC, ACFC and fusion of DGPAs with two parallel processors

Fusion of DGP as with two parallel processors	ANGER	JOY	NEUTRAL	DISGUST	FEAR	SADNESS
ANGER	**87.0**	28.0	8.0	5.0	20.0	0.01
JOY	3.0	**22.0**	1.0	0.0	6.0	5.0
NEUTRAL	10.0	50.0	**91.0**	55.0	52.0	25.0
DISGUST	0.0	0.0	0.0	**25.0**	1.0	10.0
FEAR	0.0	0.0	0.0	5.01	**21.0**	15.0
SADNESS	0.0	0.0	0.0	0.0	0.0	**35.0**

Table 5 indicates that he highest rate of learning occurs in the neutral speech state with the identification rate of 91%.

Table 6 shows the rate emotional speech recognition by using the combination of MFCC, ACFC and fusion of DGPAs classification method with two parallel processors.

Table 6. Speech emotion recognition using the combination of MFCC, ACFC and fusion of DGPAs with three parallel processors

Fusion of DGP as with three parallel processors	ANGER	JOY	NEUTRAL	DISGUST	FEAR	SADNESS
ANGER	**89.00**	15.0	7.10	5.0	0.0	10.0
JOY	0.0	**8.0**	0.0	5.0	0.0	15.0
NEUTRAL	11.00	62.0	**92.90**	50.0	60.0	24.3
DISGUST	0.0	0.0	0.0	**20.0**	9.0	10.0
FEAR	0.0	15.0	0.0	10.0	**20.50**	15.0
SADNESS	0.0	0.0	0.0	10.0	50.10	**25.7**

According to table 6 the highest rate of emotional speech recognition belong to the proposed model in neutral emotion by 92.90%. Table 7 indicates speech emotion recognition using the combination MFCC, ACFC to extract the features and three layers MLP methods to classify the feature's vector [14].

Table 7. Speech emotion recognition using the combination of MFCC, ACFC and three layer MLP

Fusion of DGP as with three parallel processors	ANGER	JOY	NEUTRAL	DISGUST	FEAR	SADNESS
ANGER	**85.0**	-	-	-	-	-
JOY	-	**30.0**	-	-	-	-
NEUTRAL	-	-	**87.50**	-	-	-
DISGUST	-	-	-	**25.0**	-	-
FEAR	-	-	-	-	**30.0**	-
SADNESS	-	-	-	-	-	**10.0**

According to table 7 the highest rate of emotional speech recognition by MLP classification belongs to neutral. But by comparing table 6 and table 7, the highest rate of speech emotion recognition belongs to the proposed model.

6. Conclusions

This paper introduces a new classification method based on multi- constraint partitioning by learning automata on emotional speech signals.

We have used six emotional states (anger, joy, neutral, disgust, fear and sadness), on the Berlin dataset, and the simulation environment has been MATLAB 2014. The proposed model consists of two main parts: feature extraction and classification. In feature extraction part, the proposed model used the combination of MFCC and ACFC models. In the part of classification multi-constraint partitioning, different type of learning automata are used including variable structure, fixed structure learning automata and DGPA.

According to experimental results, each model of classification has some disadvantages such as poor learning speech and low performance. Therefore we introduced a fusion model of learning automata with different number of parallel processors. Experimental result shows that the combination of MFCC, ACFC and fusion of DGPAs with three parallel processors has a higher performance on emotional speech signals than other methods.

Also, by comparison between tables 1 to 6 we have found that the highest rate of speech emotion recognition belong to neutral emotion. Although the proposed model have been compared by MLP model.

References

[1] R´azuri, J.G., D. Sundgren, R. Rahmani, A. Larsson, A.M. Cardenas, and I. Bonet. "Speech emotion recognition in emotional feedback for human - robot interaction ". International Journal of Advanced Research in Artificial Intelligence, vol. 4, pp. 20-27, 2015.

[2] Ayadi, M., S. Kamel, and F. Karray. "Survey on speech emotion recognition: features, classification schemes and databases". Pattern Recognition, vol. 443, pp. 572- 587, 2011.

[3] Cowie, R., S. Tspatsoulis, S. Kollias, and J. Taylor. "Emotion recognition in human-computer interaction". IEEE signal Processing, vol. 18, pp. 32-80, 2001.

[4] Seehapoch, T. and S. Wongathanavasu. "Speech emotion recognition using support vector machines". 5th International Conference on Knowledge and Smart Technology (KST), 2013, pp. 621 - 625.

[5] Ververidis, D. and C. Kotropoulos. "Emotional speech recognition: resources, features and methods". Elsevier Speech communication, vol. 489, pp. 1162- 1181, 2006.

[6] Hozjan, V. and Z. Kacic. "Context-independent multilingual emotion recognition from speech signal". Int. J. Speech Technol, vol. 6, pp. 311-320, 2003.

[7] cahn, J. "The generation of affect in synthesized speech". Journal of the American Voice I/O Society, pp. 1-19, 1990.

[8] Zhang, Q., N. An, K. Wang, F. Ren, and L. Li. "Speech Emotion Recognition using Combination of Features". Forth International Conference on Intelligent control and Information Processing (ICICIP), 2013.

[9] Bojanic, M., V. Crnojevic, and V. Deliv. "Application of neural network in emotional speech recognition". IEEE, pp. 20-22, 2012.

[10] Price, J. and A. Eydgahi. "Design an automatic speech recognition system using Matlab". 9th International Conference on Engineering Education, 2006, pp. 100-106.

[11] Adell, J., A. Bonafonte, and D. Escudero. "Analysis of prosodic features: towards modelling of emotional and pragmatic attributes of speech". Sociedad Española para el Procesamiento del Lenguaje Natural, vol. 35, pp. 277- 284, 2005.

[12] Wu, D., T.D. Parsons, and S.S. Narayanan. "Acoustic feature analysis in speech emotion primitives estimation". Interspeech, pp. 785-788, 2010.

[13] Lotfi, E. "Mathematical modeling of emotional brain for classification problems". Proceedings of IAM, vol. 21, pp. 60- 71, 2013.

[14] Muda, L., M. Begam, and Elamvazuthi. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW techniques)". Journal of Computing, vol. 23, pp. 138-143, 2010.

[15] Thathachar, M.A.L. and P.S. Sastry. "Varieties of learning automata". An Overview in IEEE Transaction on System, vol. 326, pp. 711-722, 2002.

[16] Narendra, K.S. and M.A.L. Thathachar. "Learning automata". An Introduction in Prentice Hall, 1974.

[17] Narendra, K.S. and M.A.L. Thathachar. "Learning automata - A survey". IEEE Transactions on Systems, Man and Cybernetics, vol. 44, pp. 323-334, 1974.

[18] Horn, G. and B.J. Oommen. "Solving Multiconstraint Assignment Problems Using Learning Automata". IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 401, pp. 6 - 18, 2009.

[19] Eyben, F., M. Wöllmer, and B. Schuller. "Opensmile: the munich versatile and fast open source audio feature extractor". 10 Proceedings of the International Conference on Multimedia, 2010, pp. 1459-1462.

[20] Harimi, A., A. Shahzadi, A.R. A. R. Ahmadyfard, and K. Yaghmaie. "Classification of emotional speech spectral pattern features". Journal of AI and Data Mining, vol. 21, pp. 53-61, 2014.

[21] Burkhardt, F., A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss. "A database of German emotional speech". Interspeech, vol. 5, pp. 1517-1520, 2005.

Sara Motamed She is a Ph.D candidate of artificial intelligence engineering in Science and Research University of Tehran, Iran. She is a lecturer at Islamic Azad University of Fouman, Iran. Her interests are cognitive science image processing, machine vision, signal processing, and etc. She is reviewer of Journal of Information Systems and Telecommunication (JIST).

Saeed Setayeshi Associate Professor at Amirkabir University of technology, Tehran, Iran. His research interests include medical imaging systems, neural networks and fuzzy control, Medical radiation instrumentation, etc. He has presented and published many articles in scientific journals and conferences.

Azam Rabiee Assistant Professor from 2012, and lecturer from 2004 at Islamic Azad University, Dolatabad Branch, Isfahan, Iran.

Formerly, she was with Computational NeuroSystems Lab., Brain Science Research Center, KAIST in Daejeon, Korea from 2010 to 2011, as a visiting researcher. Her research interest includes speech processing, machine learning and biologically-inspired artificial intelligence approaches.

Arash Sharifi Received the B.Sc degree in computer hardware engineering from Islamic Azad university South Tehran branch, the M.Sc and Ph.D degree in computer artificial intelligence engineering from Islamic Azad University, Science and Research branch (SRBIAU), Tehran, in 2004, 2006, and 2010, respectively. He is currently assistant professor and the head of Department of Computer artificial intelligence engineering in SRBIAU University. His current research interests include machine learning, neural networks, deep learning and evolutionary algorithms.